Clustering Techniques

Oct 09 4:00- 6:00 PM

Soft — GMM

&

Hard —

K-means

Manu K. Gupta

Department of Management Studies,

MFS of data science and AI,

IIT Roorkee.

Machine Learning Module

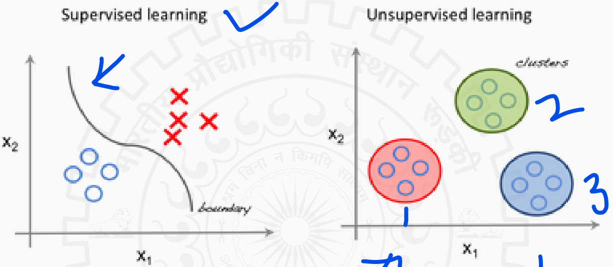
Outline

Clustering

K-mean Clustering

Agglomerative hierarchical clustering

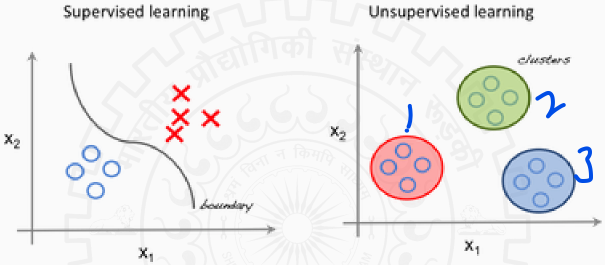
Clustering


Supervised learning
x_2
x_1
boundary
Unsupervised learning
clusters
x_2
x_1
1
2
3

labels

X

Alp.



To get an intuition about the structure of the data.

- Understanding different customer groups
- Understanding similar companies.
- Segmenting customers
- Bio-informatics

- Clusters and clustering.

- Clusters and clustering.
- Data often form groups that are close to each other, so called clusters.
- Process of discovering such structures is called clustering.

Clustering algorithms classification

Clustering algorithms classification

- Hard or crisp clustering
  - Each point is assigned to exactly one cluster.
- Soft or fuzzy clustering
  - Each point is assigned to several clusters with certain probability that add upto 1.

Clustering algorithms classification

- Hard or crisp clustering
  - Each point is assigned to exactly one cluster.
- Soft or fuzzy clustering
  - Each point is assigned to several clusters with certain probability that add upto 1.

K-mean clustering is hard clustering.

K-mean Clustering

K-mean clustering

Step 1 We initialize k points.

Step 2 Categorize each item to its closest mean and update mean's coordinate.

Step 3 Repeat the process unless stopping criterion is met.

Step 4 Report the clusters.

K-mean clustering

K = 4

Step 1 We initialize k points.

Step 2 Categorize each item to its closest mean and update mean's coordinate.

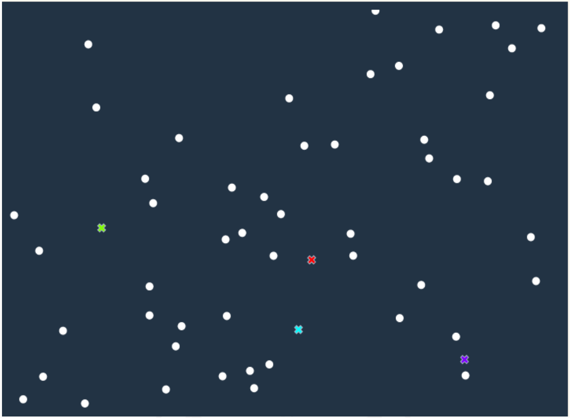
Step 3 Repeat the process unless stopping criterion is met.

Step 4 Report the clusters.

STEPS = 100

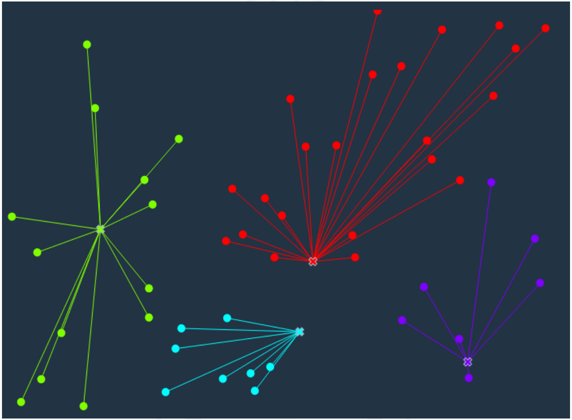
- Stopping Criterion: Maximum number of steps or convergence
- Distance: Euclidean, Manhattan

Visualization: K-mean clustering K = 4 and N = 50



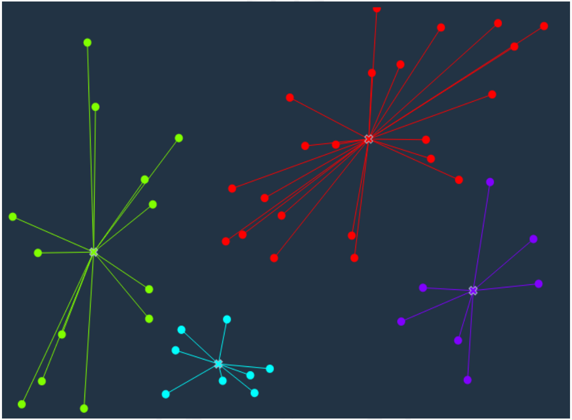
Iteration 0

Visualization: K-mean clustering $K = 4$ and $N = 50$



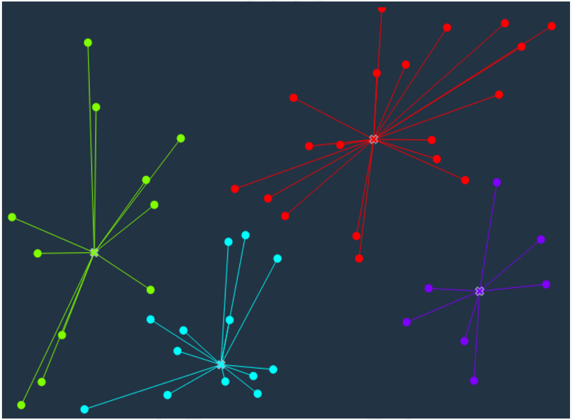
Iteration 1

Visualization: K-mean clustering K = 4 and N = 50



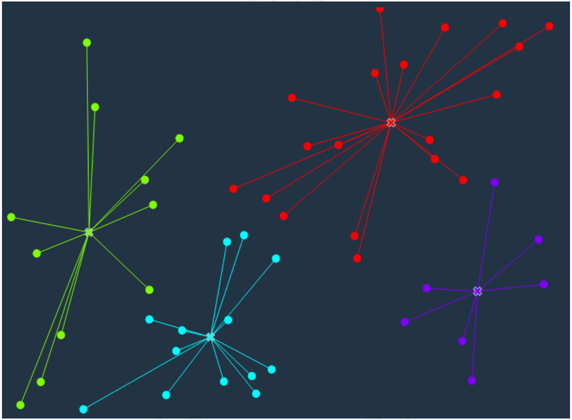
Iteration 2

Visualization: K-mean clustering K = 4 and N = 50



Iteration 3

Visualization: K-mean clustering K = 4 and N = 50

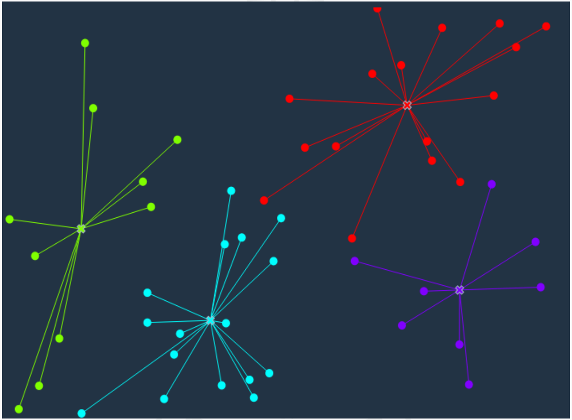


Iteration 4

Visualization: K-mean clustering K = 4 and N = 50

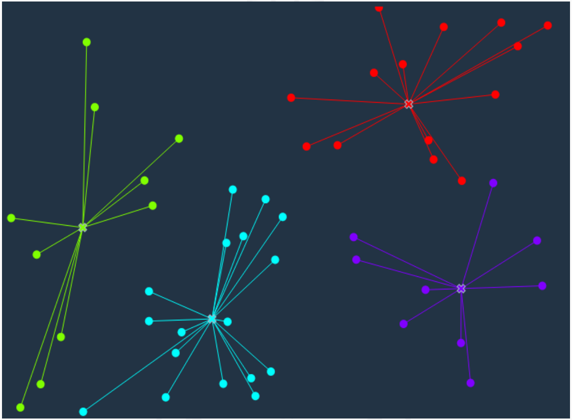
Iteration 5

Visualization: K-mean clustering K = 4 and N = 50



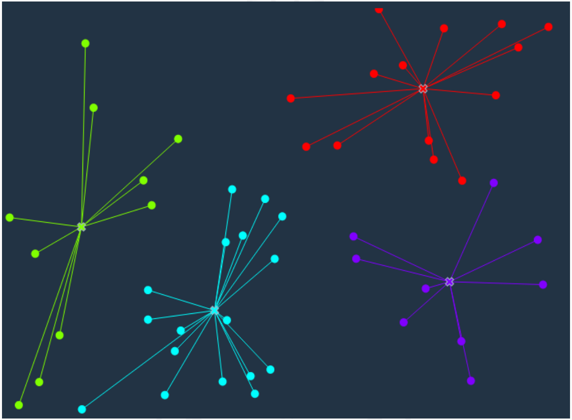
Iteration 6

Visualization: K-mean clustering K = 4 and N = 50



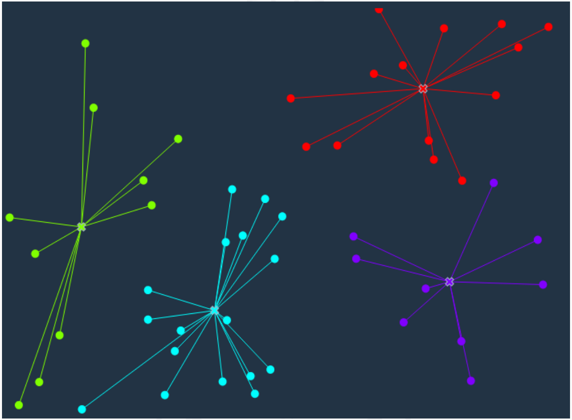
Iteration 7

Visualization: K-mean clustering K = 4 and N = 50



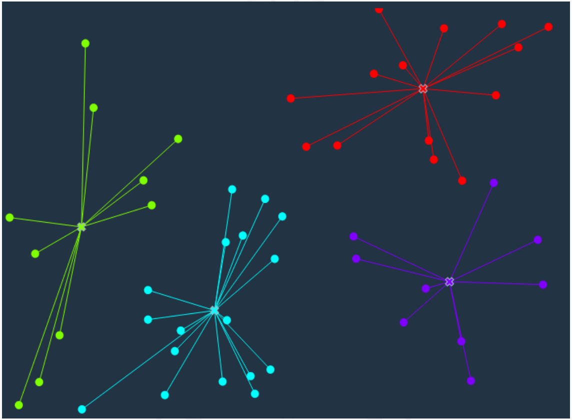
Iteration 8

Visualization: K-mean clustering K = 4 and N = 50



Iteration 9

Visualization: K-mean clustering K = 4 and N = 50



Iteration 10

Toy example

We want to group the visitors to a website using their age:

$$X = \{15, 16, 17, 20, 21, 22, 25, 36\}$$

Let's say $K = 2$. Distance of ith element: $|x_i - c_i|$

C1 = 16 and C2 = 25

Iteration 1:

Cluster 1 = {15, 16, 17, 20}

Cluster 2 = {21, 22, 25, 36}

C1 = 17

C2 = 26

Iteration 2

Cluster 1 = {15, 16, 17, 20, 21}

Cluster 2 = {22, 25, 36}

Major challenge in K-mean clustering?

Major challenge in K-mean clustering?

- How to choose K?

Major challenge in K-mean clustering?

- How to choose K?

Typically, there are two methods:

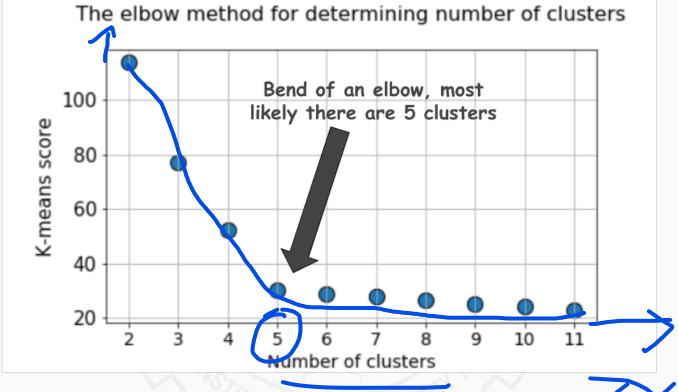
Major challenge in K-mean clustering?

- How to choose K?

Typically, there are two methods:

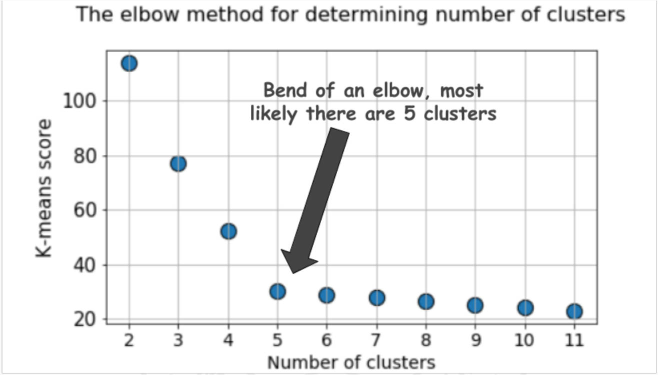
- Elbow method
- DB (Davis-Bouldin) index

Elbow method



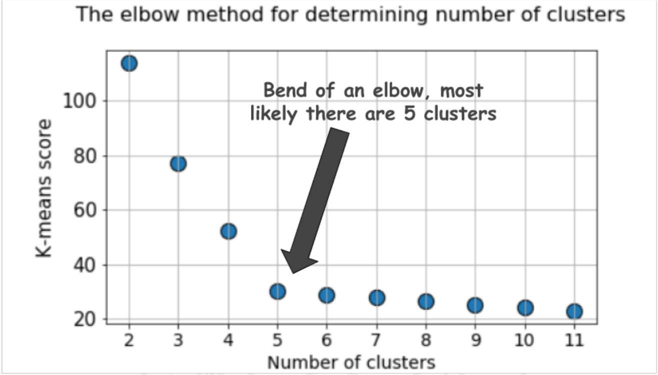
SSE

Elbow method



- How do I find out K-means score?

Elbow method



- How do I find out K-means score?
- Within groups SSE.

DB index

To choose K or the number of clusters

- Cluster dispersion:

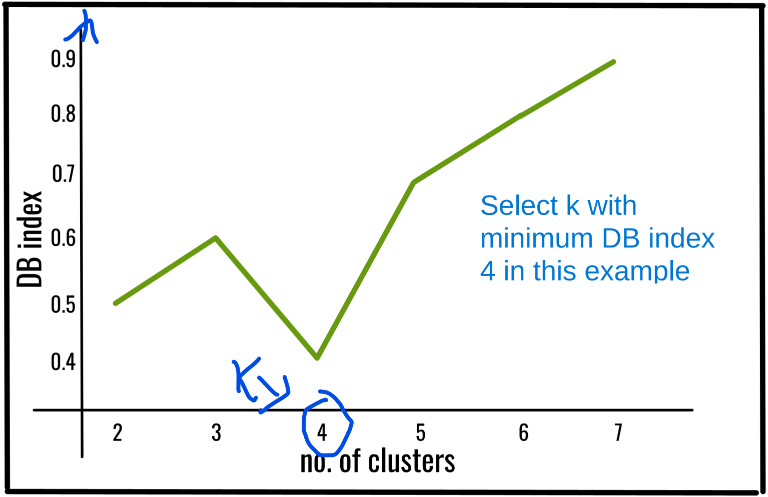
$$\delta_k := \sqrt{\frac{1}{N_k} \sum_{n \in \mathcal{C}_k} ||x_n - c_k||^2}$$

- Cluster similarity of two clusters:

$$s_{kl} := \frac{\delta_k + \delta_l}{||c_k - c_l||}$$

- DB index

$$\boxed{V_{DB} := \frac{1}{K} \sum_{k=1}^{K} \max_{l \neq k} S_{kl}}$$

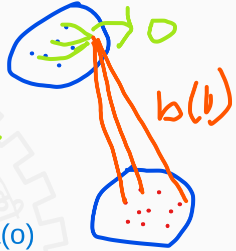

DB index
0.9
0.8
0.7
0.6
0.5
0.4
2
3
4
5
6
7
no. of clusters
Select k with
minimum DB index
4 in this example

- Did we choose the parameters (e.g., number of clusters) in the most optimal way?

Classification --- Accuracy/Recall
Regression --- R2 score/MSE/RMSE
Clustering --- ??

- Did we choose the parameters (e.g., number of clusters) in the most optimal way?

Silhouette score.



Silhouette score for observation o:

$$\boxed{S(o) = \frac{b(o) - a(o)}{\max\{b(o), a(o)\}}}$$

b(o) a(o)

- $a(o)$ is the average distance to other samples within cluster.
- $b(o)$ is the average distance to other samples in other clusters.

Case 1: b(0) > a(o): $= \frac{b(o) - a(o)}{b(o)}$

$\Rightarrow$ Close to 1.

Explanation

Case: b(o) is similar to a(o): S(o) will be close to zero

Overlapping clusters

Case 3: b(o) < a(o):

$$s(o) = \frac{b(o) - a(o)}{a(o)} = \underbrace{\frac{b(o)}{a(o)}}_{<1} - 1$$

$$< 0$$

Not desirable...

Explanation

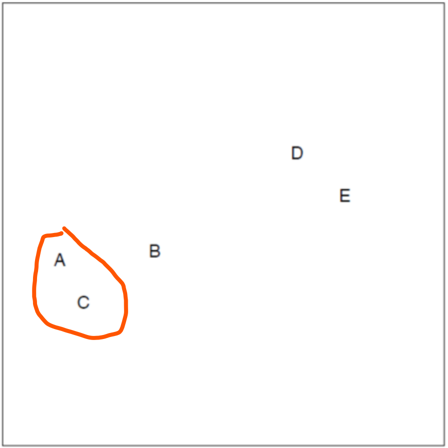
Global silhouette coefficient: average of the sum of $S(o)$ for each point

Demo

- K-mean clustering.
    - pre-specify the number of clusters K.
    - Depends on the random initialization.
- Hierarchical clustering is an alternative approach
    - Agglomerative (fine to coarse) hierarchical clustering.
    - Devisive (coarse to fine) hierarchical clustering

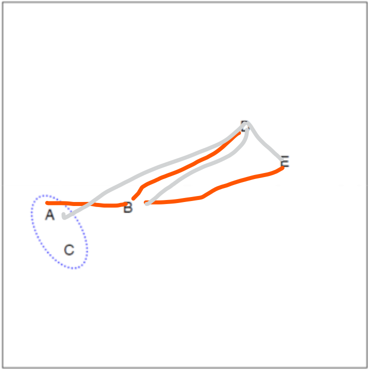
K means clustering:
How to decide the value of K?
-- DB index
-- Elbow method
-- Model evaluation

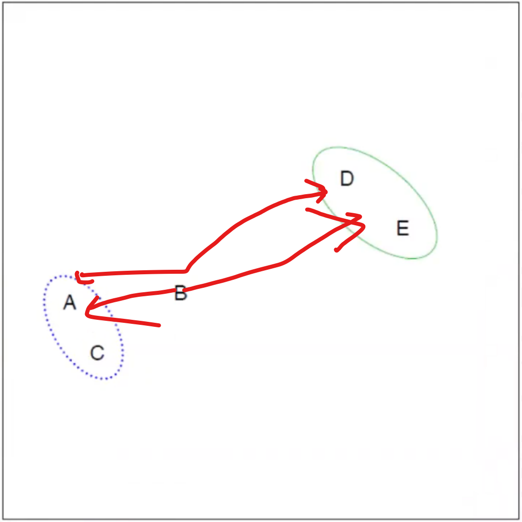
Agglomerative hierarchical clustering



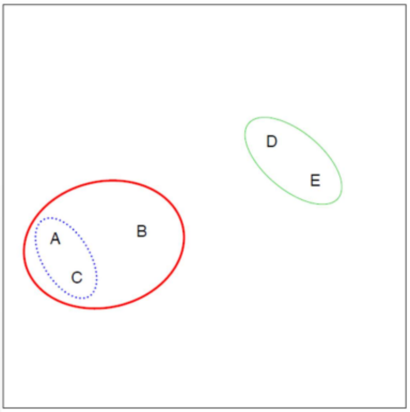
	A	B	C	D	E
A	0				
B	d_1	0			
C	d_2	d_3	0		
D	⋮	⋮	⋮		
E					

$\min\{d_{ij}\} = d_2 \rightarrow AC$

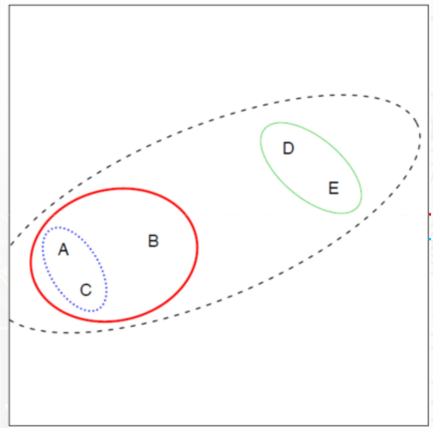

A
B
C
E

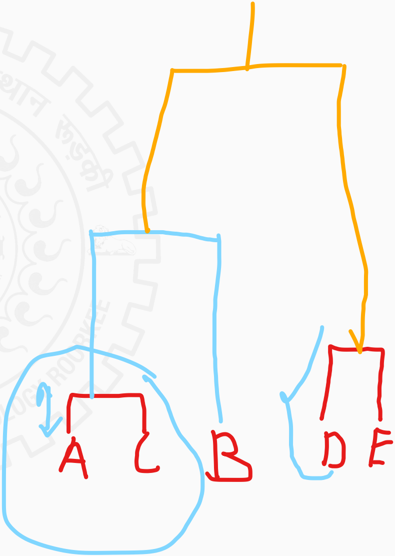


Choose the clusters based on smallest distance....

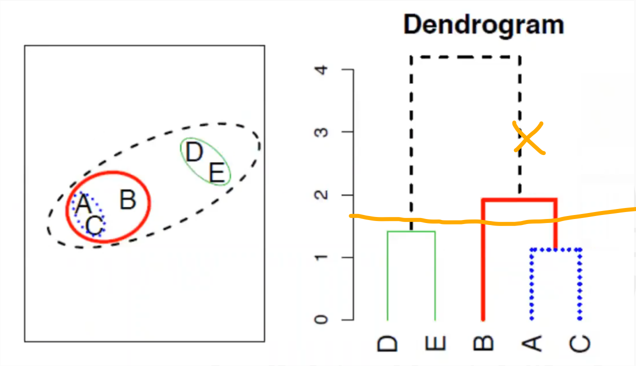

D
E
A
B
C

Dendogram





Dendrogram



C1 = {D, E}

C2 = A B C

y-axis on dendrogram is (proportional to) the distance between the clusters that got merged at that step

C1 = D, E   C2 = B   C3 = A,C

Steps in Algorithm

Step 1 Start with each point in its own cluster.

Step 2 Identify the two closest clusters. Merge them.

Step 3 Repeat until all points are in a single cluster.

What is the challenge?

What is the challenge?

How to measure the distance between clusters?

What is the challenge?

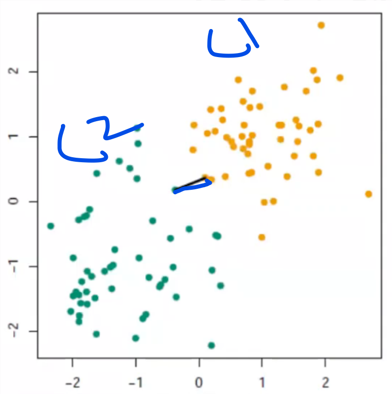
How to measure the distance between clusters?

- Single linkage: Minimal inter-cluster dissimilarity.
- Complete linkage: Maximal inter-cluster dissimilarity.
- Average linkage: Mean inter-cluster dissimilarity.
- Centroid linkage: Centroid inter-cluster dissimilarity.

Single Linkage

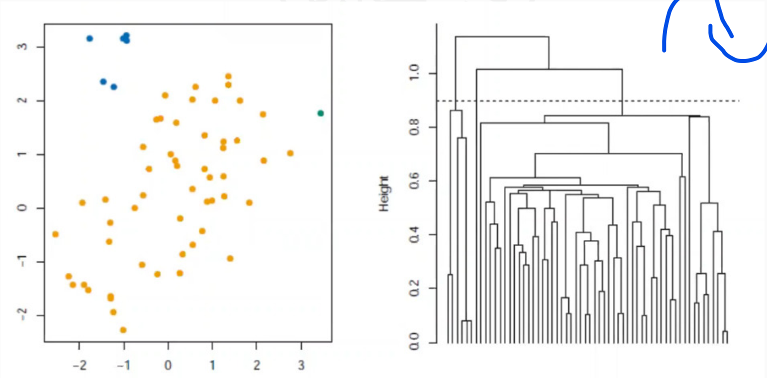
The distance between G, H is the smallest distance between two points in different groups.

$$d_{single}(G, H) = \min_{i \in G,\ j \in H} d(x_i, x_j)$$



Single Linkage

Here $n = 60, x_i = \mathbb{R}^2, d_{ij} = ||x_i - x_j||_2$. Cutting the tree at $h = 0.9$ gives the clustering assignments marked by colors.

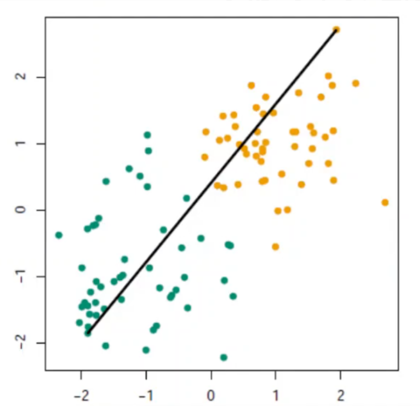


> Cut interpretation: For each x_i, there is another point x_j in its cluster such that $d(x_i, x_j) \leq 0.9$

Complete linkage

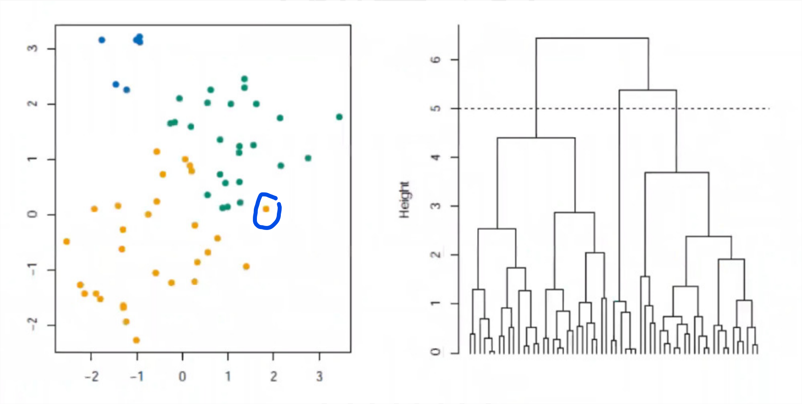
The distance between G, H is the largest distance between two points in different groups.

$$d_{complete}(G,H) = \max_{i \in G,\ j \in H} d(x_i, x_j)$$



Complete Linkage

Here $n = 60, x_i = \mathbb{R}^2, d_{ij} = ||x_i - x_j||_2$. Cutting the tree at $h = 5$ gives the clustering assignments marked by colors.



Cut interpretation: For each x_i, every other x_j in its cluster satisfies that $d(x_i, x_j) \leq 5$

Single Vs Complete Linkage

- Single linkage suffers from *chaining*.

Single Vs Complete Linkage

- Single linkage suffers from *chaining*.
    - poorly separated, distinct clusters are merged at an early stage

Single Vs Complete Linkage

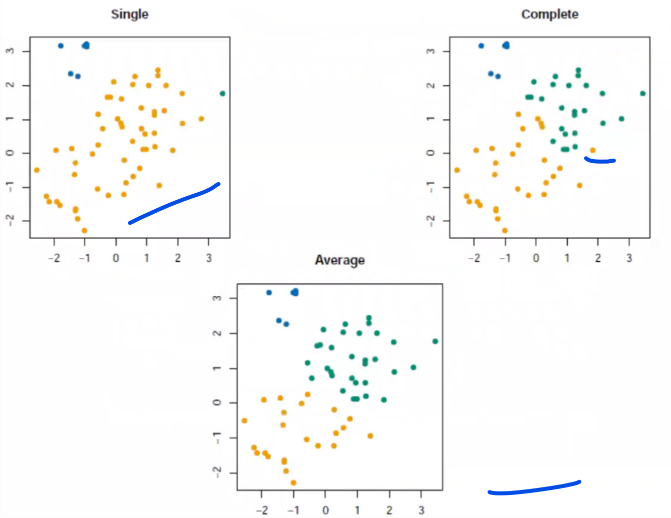
- Single linkage suffers from *chaining*.
    - poorly separated, distinct clusters are merged at an early stage
- Complete linkage avoids chaining but suffers from *crowding*.

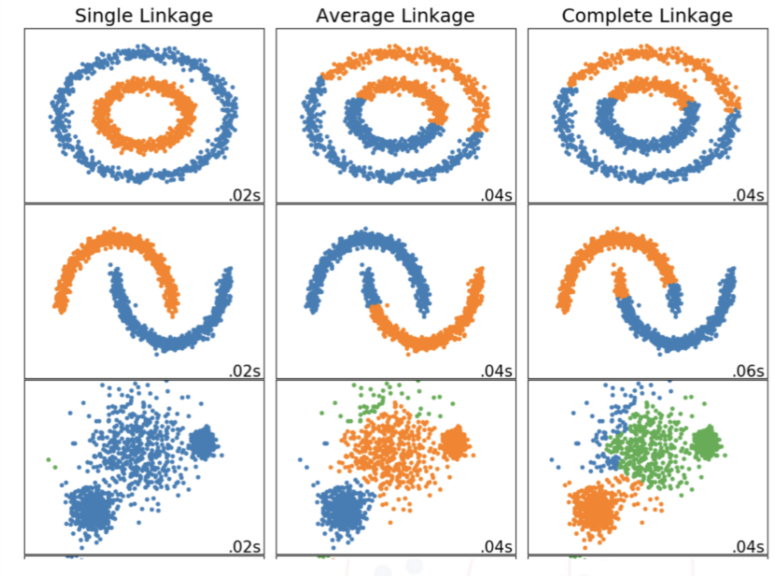
Single Vs Complete Linkage

- Single linkage suffers from *chaining*.
    - poorly separated, distinct clusters are merged at an early stage
- Complete linkage avoids chaining but suffers from *crowding*.
    - A point can be closer to points in other clusters than to points in its own cluster.

Single Vs Complete Linkage

- Single linkage suffers from *chaining*.
    - poorly separated, distinct clusters are merged at an early stage
- Complete linkage avoids chaining but suffers from *crowding*.
    - A point can be closer to points in other clusters than to points in its own cluster.
- Average linkage tries to strike a balance.


Single
Complete
Average
3
2
1
0
−1
−2
−2 −1 0 1 2 3



Single Linkage
Average Linkage
Complete Linkage
.02s
.04s
.04s
.02s
.04s
.06s
.02s
.04s
.04s

Denodogram

useful in identifying
the number of clusters
using horizontal lines

demo

Summary:

unsupervised learning
Labels are not known
Clusters and clustering
Hard vs soft
K-means clustering
Model Evaluation
Chaining property is missing in k-means
AC by using single linkage
other linkages --- complete, centroid and average

Thank you!

manu.gupta@ms.iitr.ac.in